Development of a Novel Genetic Algorithm Search Method (GAP1.0) for Exploring Peptide Conformational Space

A. Y. JIN, F. Y. LEUNG, D. F. WEAVER*

Department of Chemistry, Queen's University, Kingston, Ontario, K7L 3N6, Canada

Received 20 January 1997; accepted 29 July 1997

ABSTRACT: A genetic algorithm-driven search method (GAP1.0; Genetic Algorithm Peptide search, version 1.0) has been developed for the computational exploration of peptide conformational space. The suitability of a variety of genetic algorithm operators was evaluated through representative calculations on the pentapeptide [Met]-enkephalin (Tyr–Gly–Gly–Phe–Met). GAP1.0 was successful in efficiently elucidating backbone conformational features observed in the global minimum energy structure. Furthermore, the program readily identified the tremendous diversity among [Met]-enkephalin conformers under physiological conditions. It is concluded that GAP1.0 provides a useful extension to the current repertoire of conformational analysis techniques. © 1997 John Wiley & Sons, Inc. *J Comput Chem* **18**: 1971–1984, 1997

Keywords: genetic algorithm; peptide; conformer; enkephalin; search

Introduction

The conformational flexibility of peptides plays a dominant role in determining bioactivity; for example, the ability of a small peptide to assume many conformations may lead to interactions at multiple receptor sites.[1,2] Accordingly, theoretical and experimental techniques of peptide conformational analysis are important tools in the rational design of peptides and peptidomimetics. Recent computer advances have enhanced theoretical conformational studies of peptides and other biomolecules via the empirical force field approach.[3] Unfortunately, in the absence of crytstallographic or spectroscopic experimental data, a definitive investigation of molecular structure is precluded by a ''multiple-minima problem,''[4] originating from the large number of variables required to describe conformations within the force field context. Even for small molecules, the complexity of the force field expression defines a pro-

* Also affiliated with the Department of Neurology, Queen's University.

Correspondence to: D. F. Weaver, e-mail: weaverd@quchem.queensu.ca

Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada

CCC 0192-8651 / 97 / 161971-14

foundly labyrinthine potential energy hypersurface (''conformational space''), thus generating a vast pool of potential energy minima from which many may have sufficiently low energy to be biologically significant. Current computational methods for searching conformational space tend to be inefficient at locating multiple potentially bioactive conformations. As such, new efficient search algorithms are needed.

Current search methods fulfill either an exploratory or an exploitative role. Exploratory approaches have a random nature, as exemplified by Monte Carlo (MC) search techniques.[5] This approach offers a sampling of widely separated points in a molecule's conformational space through a random selection of conformers. Often, the MC method is hybridized with energy minimization routines to insure that unreasonably high energy conformers are ''filtered'' from the output.[6] Exploitative approaches utilize the information presented in a conformer's force field expression. For example, in a molecular dynamics (MD) trajectory, the positions of corresponding nuclei in consecutively generated conformers are related through an application of Newton's second law of motion[7]; accordingly, for a given nucleus, its position, velocity, and the net force acting on it determine its new position in a subsequently generated conformation. Because the MD approach searches a small region of conformation space thoroughly, it is particularly useful in conjunction with experimentally derived structural data.[8]

To derive new search procedures, natural processes have provided metaphors for computational search algorithms.[9] Genetic algorithms and simulated annealing—modeled on evolutionary and thermodynamic phenomena, respectively—have received much attention as new classes of stochastic search methods.[10] Although simulated annealing has received wide attention in peptide conformational analysis, genetic algorithms have not. The aim of this study was to explore the capability of genetic algorithms (GAs) to yield novel insights on peptide conformational space. Although the foundation of GA theory was outlined two decades ago,[11] the application of GAs in chemistry is only now emerging. A diverse range of topics, from chemometrics[12] to the modeling of drug–receptor docking,[13] have been examined using GA search methods. GAs have shown preliminary promise as a conformational search technique for both large and small molecules.[14] This success may be attributed to the incorporation of both exploratory and exploitative mechanisms within a typical GA. This permits a very flexible exploration/exploitation dichotomy which can adapt to the current search conditions. Although these results are encouraging, several important issues remain unresolved. For example, an *a priori* means of establishing the optimal set of GA parameters has not been established. In addition, because GAs were initially implemented as a global optimization technique, they must be modified for conformational search tasks in which one wishes to find many different low energy conformers. Finally, the complex nature of peptide folding processes is likely to confound a simple GA-based conformational search.

This study reports the development of a program called GAP1.0 (Genetic Algorithm Peptide search, version 1). The modification of the GAP1.0 conformational space search by various parameter settings has been investigated through representative calculations on the endogenous opioid peptide, [Met]-enkephalin (Tyr–Gly–Gly–Phe–Met). [Met]-enkephalin is an extremely important neuropeptide central to diverse phenomena extending from pain management and drug addiction to cough suppression. The bioactivity of [Met]-enkephalin may arise from various distinct low energy conformers and not from the single lowest energy structure; for example, [Met]-enkephalin's inherent conformational flexibility leads to affinity for at least the μ- and δ-opioid receptor sites. GAP1.0 has been specifically designed to investigate conformational diversity at low energies.

Methods

GAP1.0 ALGORITHM PRINCIPLES

The principles of GAs were first elaborated by Holland.[11] By implementing data-manipulating operations to mimic evolutionary processes, Holland devised a procedure—a ''genetic algorithm'' —which exhibited similarities to adaptive processes found in natural and artificial settings. In developing GAP1.0, we have applied the principles of adaptation theory to peptide conformational analysis. Conceptually, this has been achieved as follows: a ''gene'' represents a torsional angle within a peptide; a peptide with a distinct conformation defines a ''parent'' (different parents have the same amino acid sequence, but different conformations); parents are ''mated'' by ''recombining genes'' to create ''offspring'' (i.e., the offspring's conformation may be different from

either of its parent' conformations); the ''fitness'' of an offspring reflects its conformational energy as determined with an empirical force field calculation (low energy conformers become ''survivors,'' whereas high energy conformers are ''evolutionarily unsuccessful''); to insure ''survival of the fittest'' low energy peptide conformers are favored for mating with other population members, thereby hopefully producing offspring that are also low energy conformers; to insure favorable genetic variability, ''mutations'' are randomly imposed on the mating process. Procedurally, gene values (i.e., peptide torsional angles) are encoded as binary numbers; a ''chromosome'' (a collection of torsional angles) is thus a binary string. To manipulate these binary strings, a number of adaptation operators have been put in place: a ''crossover operator'' determines how genes are recombined to generate offspring; a ''selection operator'' identifies population members that are fit to function as parents in subsequent generations; a ''mutation operator'' randomly flips bits in the binary string chromosome to either 0 or 1, with a probability defined by the ''mutation rate.''

The success of this algorithm is founded on the relationship between performance and ''schema,'' or ''building blocks'' within the chromosome. In a binary chromosome representation, schema consist of patterns of bits. For example, ''1×01'' is a schemata in which the first, third, and fourth bit positions must have the values ''1,'' ''0,'' and ''1'' respectively; the second position can be occupied by either ''0'' or ''1.'' The ''order'' of a schemata is given by its number of well-defined (i.e., ''0'' or ''1'') positions and the ''defining length'' by its total number of positions. Therefore, the schemata ''1×01'' has an order of three and a defining length of four. After successive generations, schema that improve fitness propagate throughout the population whereas schema that diminish fitness are eliminated. This principle is known as ''the schema theorem'' and is responsible for the suitability of GAs as an optimization method.

STRUCTURE OF GAP1.0 PROGRAM

The GAP1.0 code was organized as shown in Figure 1. For input, the program required values for the population size, the mutation rate, the number of generations, a seed value required by a random number generator, and the necessary information for energy calculations using the ECEPP/2 force field (e.g., primary residue sequence, cystine disulfide bond data).[3d] These parameters were read in the subroutine INITDAT. For a given primary sequence, peptide conformations were randomly generated in the subroutine INITPOP. The initial population size (i.e., the initial set of parents) was half the full population size (i.e., total number of parents and offspring) specified by the user. Conformations were generated by assigning random values to all phi, psi, and chi peptide backbone angles; all omega dihedral angles were set to 180° throughout the GA run. Each angle was represented by a string of eight bits, thereby allowing initial values that were multiples of $360/255 \cong 1.41°$; angles in subsequent generations were rounded to the nearest integer. The torsional angle data for each conformer was then translated into an ECEPP input file. Within subroutine EVALUATE, the energy of each conformer was calculated through a subroutine call to the ECEPP program. Based on energy, the subroutine FITCONF assigned fitness values to each conformer using the fitness-ranking method.[10] Fitness ranking permitted an even distribution of fitness values (from 0 to 0.99) throughout the entire initial population and every parent population thereafter, regardless of the energy of each conformer. For each generation, the highest, lowest, and average population energies were recorded by the subroutine RESULTS. Additionally, the conformations present in the initial generation and those corresponding to 1/4, 1/2, 3/4, and full completion of the GA run were recorded by the subroutine RE-

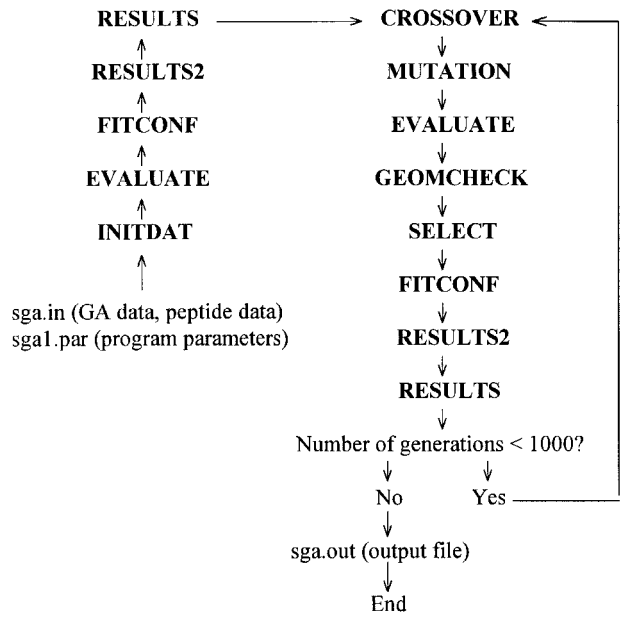

FIGURE 1. Structure of GAP1.0 source code.

SULTS2. Once the energies and fitness values for the initial parent population were evaluated, the population was permitted to ''evolve.'' The offspring were generated through the subroutine CROSSOVER until a full-size population was obtained. Within this subroutine, three different crossover schemes were implemented: one-point, two-point (chromosomes are crossed over at two points); and uniform crossover. The offspring were then perturbed in the MUTATION subroutine. The ECEPP energies of all conformers were calculated in EVALUATE and the offspring were checked against the parents for conformational similarity by the subroutine GEOMCHECK. An offspring was considered to be similar to a parent if more than half of its movable torsional angles were within 5° of the corresponding angles in a parent. If a similar parent–offspring pair was found, then the highest energy conformer of the pair was mutated at a high mutation rate. Once all of the offspring had been examined within the GEOMCHECK subroutine, and the energies had been recalculated, the selection criterion was applied by SELECT. Based on energy, conformers from the best-performing (i.e., lowest energy) half of the population were chosen to become parents for the next generation. At the completion of the GA run, summaries pertaining to energy and usage of the GEOMCHECK subroutine were generated. Also, parent conformers from the final generation were energy-minimized.

[MET]-ENKEPHALIN CALCULATIONS USING GAP1.0

Calculations were performed on IBM RS/6000 RISC workstations operating under AIX, and on Sun workstations operating under SunOS Unix. Source code for the genetic algorithm subroutines was written in the ANSI Fortran 77 Standard. ECEPP/2 and the accompanying parameter files were obtained from the QCPE.[15] Minor modifications to ECEPP/2 source code were made to facilitate use within AIX XL Fortran and to permit the passing of variables between ECEPP/2 and the GAP1.0 code. Hierarchical clustering analysis,[16] using the average link method, was performed using the ''method = average'' option within the ''hclust'' procedure in the commercial package SPLUS.[17] Fuzzy c-means clustering analysis was done using the implementation of Bezdek.[18] Energy minimization was performed using Powell's method.[6a, 19]

The molecule used for this study was the pentapeptide [Met]-enkephalin (Tyr–Gly–Gly–Phe–Met). The abundant structural data associated with this highly flexible and biologically important molecule provided reliable standards by which the performance of the GA could be evaluated.[20] The consistency of the GA program was ascertained by comparing the results across three data sets, each generated using a different random number generator seed. Within each data set, three disparate levels for both the mutation rate and the population size—resulting in nine combinations—were implemented with the intention of revealing any characteristics that were dependent on these parameters. The mutation rate values examined were 0.07, 0.05, and 0.03. At mutation rate 0.07, for example, one could expect that 0.07×8 bits/torsional angle $\times$ 24 torsional angles/conformer $\cong$ 13 bits/conformer would be flipped. The population sizes examined consisted of 50, 100, or 200 members, including both the parents and their offspring. Conformational and energy data were obtained from the ''selected populations'' only; that is, from parent populations of 25, 50, or 100 conformers. Each GA run was terminated after 1000 generations.

For each data set, the lowest and highest parent conformer energies, the average parent conformer energy, and the number of mutants arising from the GEOMCHECK subroutine were recorded for each generation. The conformations of the parents were recorded at the initial, 250th, 500th, 750th, and the final generation. Each conformer in the final parent population was also energy minimized. For each recorded set of parent conformers, the distribution of phi and psi angles for each residue was displayed in a Ramachandran-type scatterplot.[21] Additionally, two cluster analysis techniques—the hierarchical average link method[16] and the fuzzy c-means clustering algorithm[18]—were used to determine the presence of similar conformers within the parent population. The similarity measure used the Cartesian coordinates of the conformers to calculate the Euclidean distance.

Results

STRUCTURAL ANALYSIS OF [MET]-ENKEPHALIN

The conformational diversity revealed through the GA was reflected in the distributions of the phi and psi angles of each residue over the entire population as displayed in a Ramachandran-type

scatterplot.[21] Because these angles determine peptide conformation, any gross conformational features are reflected in their values. Despite the variety of parameter values used in this study, the initial random distribution of points on the phi–psi map consistently resolved into similar groups contained within a more narrow range. The location of these groups also corresponded to the torsional angle values associated with secondary structure motifs. Although similar motifs were found for all runs, the number of groups and the number of points within a group were often affected considerably by both the mutation rate and the population size. Scatterplots generated from runs using different mutation rates reflected the effects of both the mutation operator and the similarity checking routine. For population sizes of both 100 and 200, a high mutation rate lead to diffuse groupings, whereas dense groups were found at the lowest mutation rate. (At population size 50, there were too few points to reveal any trends.) However, as the mutation rate was decreased, the number of groups and/or the distance between the groups often increased. This is probably an indication of the increased use of the GEOMCHECK subroutine as the mutation rate was decreased. Small shifts in the distribution of points also became apparent when similar runs from different data sets were compared. This occurred repeatedly at each population size and mutation rate, indicating that the distribution of points from the initial population slightly biased the distribution at the final generation. Nevertheless, each data set featured similar conformational characteristics. Additional conformational trends were not revealed when the population size was increased from 100 to 200. At population size 50, there were insufficient points to adequately reveal the complete phi–psi range for each residue.

Results from a representative data set are shown in Figure 2. The backbone of Tyr1, Phe4, and Met5 existed mostly in a β-sheet conformation (see Fig. 2a).[22] Two groups of points were found for Tyr1: a small group at $30° < \varphi < 70°$, $120° < \psi < 160°$; and a larger group at $160° < \varphi < -100°$, $120° < \psi < 160°$. Points for Phe4 were mostly found in one large group at $-170° < \varphi < -110°$, $120° < \psi < 160°$. The Met5 backbone conformation was also described by two sets of points: a small group at $-170° < \varphi < -130°$, $0° < \psi < 30°$; and a much larger group at $-170° < \varphi < -110°$, $120° < \psi < 180°$. Additionally, for Met5, many points were found in the range $-160° < \varphi < -140°$, $40° < \psi < -170°$ (data not shown). For both Gly residues, a wider distribution of points ($+60° < \varphi < -100°$, $0° < \psi < -100°$) was probably due to the greater conformational flexibility afforded by the absence of a large side-chain (see Fig. 2b).

After energy minimization, scatterplots for Tyr1, Phe4, and Met5 were unchanged (see Fig. 2c), whereas plots for Gly2 and Gly3 resolved into groups indicative of the type II′ β-reverse turn (see Fig. 2d). The lowest energy conformer found in this study, and a previously reported global minimum energy structure,[6] are shown in Figure 3a and b.

SEARCH BEHAVIOR OF GAP1.0

The preconvergence tendency within the genetic algorithm was most apparent when one-point and two-point crossover schemes were employed. Regardless of the degree of selection pressure, total convergence over all torsional angles occurred well before the final generation. In general, convergence occurred much sooner with decreasing mutation rate and decreasing population size. With the use of the uniform crossover operator, converged populations occurred much later in the GA run or not at all. With the addition of the GEOMCHECK subroutine, preconvergence was entirely removed from all GA runs, at all combinations of mutation rate and population size. Therefore, the remaining data were generated using both the uniform crossover operator and the GEOMCHECK subroutine.

As coarse measures of the GA's progress, both the average parent energy and the energy of the lowest energy conformer were monitored during each run. The responses of these two energies varied considerably according to the mutation rate and population size (see Table I). For population sizes of 100 or 200, the lowest energy generally decreased with declining mutation rate (see Fig. 4). For a population size of 50, declining mutation rate was accompanied by a consistent decrease in lowest energy only within the mutation rate range 0.05 to 0.07. The effect of the mutation rate on the average parent conformer energy was also in turn affected by the population size. At population sizes of 50 or 100, the average parent energy decreased according to mutation rate in the general order 0.03 (highest energy) > 0.07 > 0.05 (see Fig. 5). At a population size of 200, the average parent energy was generally highest at mutation rate 0.07 and lowest at 0.03. Contrary to expectations, neither the lowest energy or the average parent energy were consistently improved on increasing the pop-

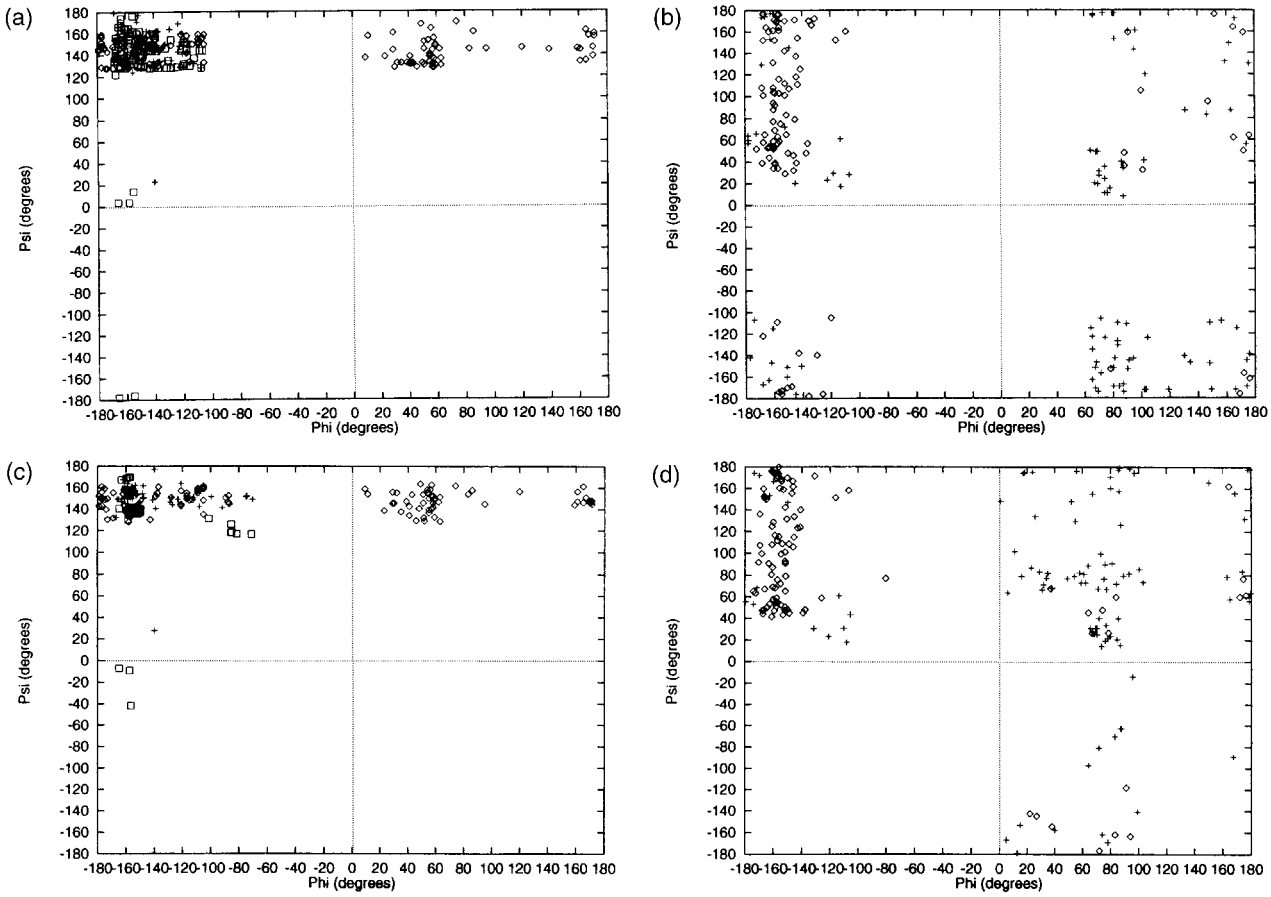

FIGURE 2. Ramachandran-type scatterplot of the phi and psi angles in: (a) Tyr1 (◇), Phe4 (+), and Met5 (□) before energy minimization of conformers in the 1000th generation; (b) Gly2 (◇) and Gly3 (+) before energy minimization of conformers in the 1000th generation; (c) Tyr1, Phe4, and Met5 after energy minimization of conformers in the 1000th generation; (d) Gly2 and Gly3 after energy minimization of conformers in the 1000th generation.

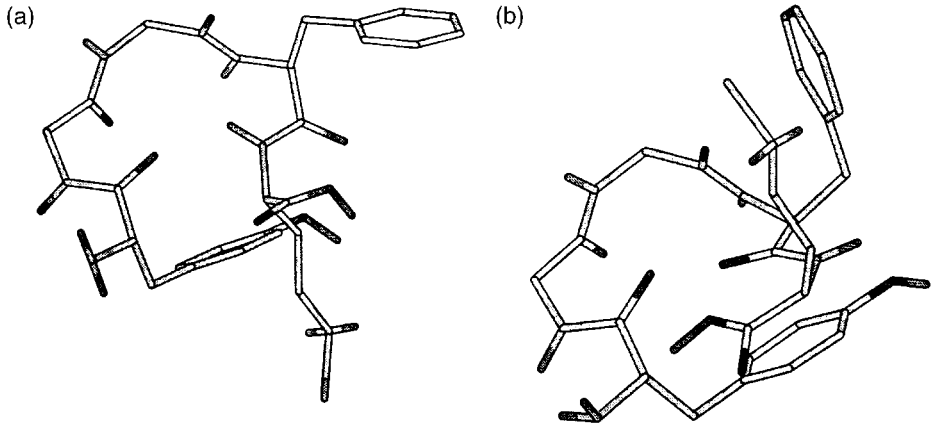

FIGURE 3. [Met]-enkephalin structures: (a) lowest energy structure found using GAP1.0 and subsequent energy minimization (−9.4 kcal / mol); (b) global minimum energy structure as reported by Nayeem et al.[6b] (−12.9 kcal / mol).

TABLE I.
Lowest Energy Conformers and Average Parent Conformer Energy from Three Data Sets.

			Before energy minimization		After energy minimization	
Data set	Population size	Mutation rate	Lowest energy (kcal / mol)	Average parent energy (kcal / mol)	Lowest energy (kcal / mol)	Average parent energy (kcal / mol)
1	200	0.07	3.3	7.5	−6.6	0.9
		0.05	2.8	6.5	−8.3	−1.2
		0.03	1.0	7.1	−5.9	0.7
	100	0.07	1.1	6.0	−4.8	0.9
		0.05	0.6	4.7	−9.1	−2.0
		0.03	0.4	6.2	−5.9	−0.9
	50	0.07	1.7	6.0	−6.7	−1.2
		0.05	1.2	4.6	−4.2	−1.3
		0.03	3.3	9.2	−6.8	−1.4
2	200	0.07	1.4	6.6	−7.6	−1.2
		0.05	0.7	7.1	−6.0	−0.7
		0.03	−2.1	6.3	−6.0	−1.5
	100	0.07	3.8	7.3	−5.3	−1.2
		0.05	−2.4	6.2	−6.6	−1.2
		0.03	1.8	8.3	−8.6	−0.9
	50	0.07	2.6	6.8	−5.0	−0.7
		0.05	1.6	4.9	−4.5	−2.2
		0.03	2.5	7.2	−4.2	−0.7
3	200	0.07	4.1	8.3	−9.4	−0.5
		0.05	0.7	7.2	−9.2	−1.1
		0.03	−1.1	6.6	−5.2	−1.7
	100	0.07	2.6	6.3	−8.3	−1.1
		0.05	0.2	5.2	−4.4	−0.8
		0.03	−0.4	6.0	−4.1	−1.5
	50	0.07	5.6	8.7	−5.0	−0.9
		0.05	2.2	6.5	−3.8	−0.8
		0.03	2.0	8.5	−4.5	−1.4

ulation size from 100 to 200 conformers. Before energy minimization was applied to the final population, the energy ranges varied in size from about 5 kcal/mol to as high as 20 kcal/mol and the average parent energy was generally over 20 kcal/mol higher than the global minimum energy reported by Nayeem et al.[6] Lowest energy conformers were found between −2.4 and 5.6 kcal/mol. After minimization, the average parent conformer energy dropped to within < 13 kcal/mol of the global energy minimum and the lowest energy conformers were found within the range −3.8 to −9.4 kcal/mol (see Table I). Neither the mutation rate nor the population size appeared to have a consistent effect on the minimized average parent conformer energy or on the lowest energy conformers.

The significance of the GEOMCHECK subroutine was reflected in the high occurrence of mutations related to its use. The frequency of use was affected strongly by the mutation rate but not by the population size. For mutation rates 0.03, 0.05, and 0.07, the percentage of the total population that was mutated by GEOMCHECK increased to maximum levels of 35%, 25%, and 15%, respectively, within the first 500 generations. This indicated a relationship between the mutation rate and the strong preconvergence tendency within the GA. Most of the mutants throughout the GA run ($> 95\%$) were generated among the offspring.

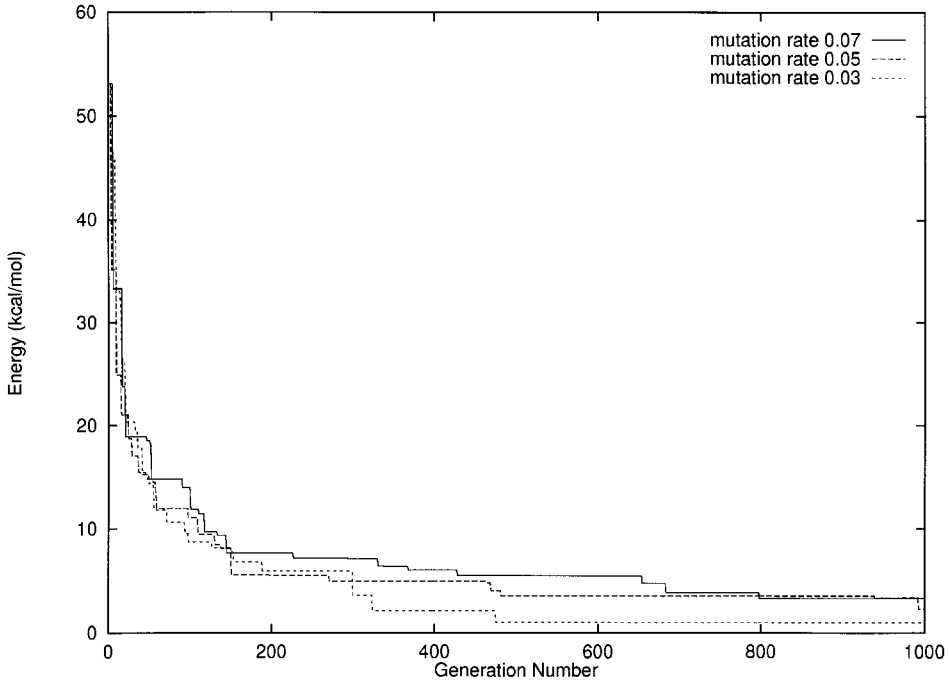

FIGURE 4. Improvement of lowest energy conformer over 1000 generations (for 100 total population members).

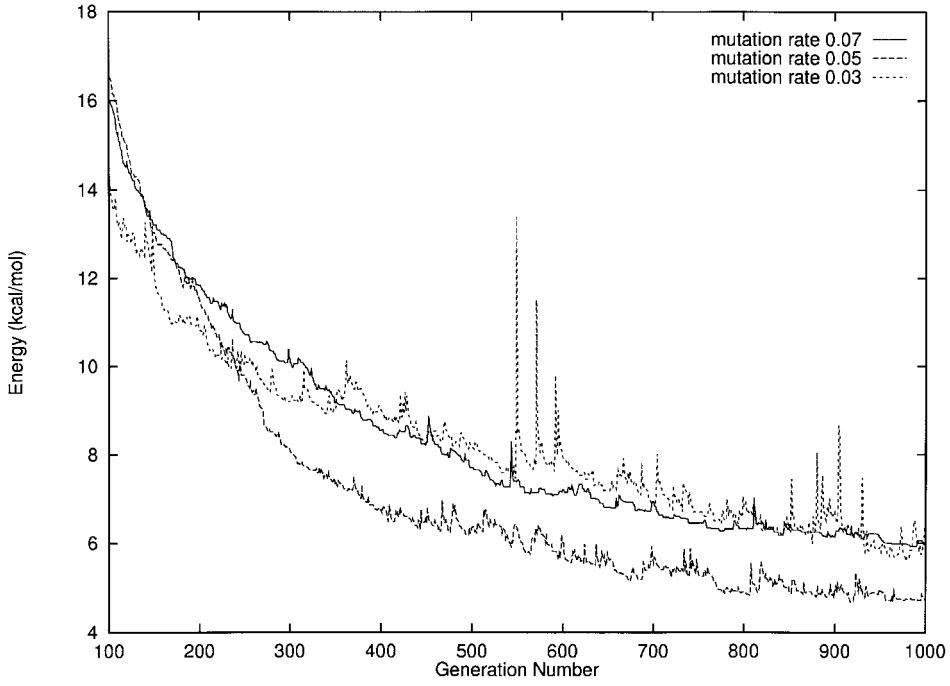

FIGURE 5. Improvement of average parent energy over 1000 generations (for 200 total population members). Data between 0 and 100 generations have been omitted to permit the use of a convenient scale in the ''Energy'' axis.

Clusters of similar conformers were not observed. Using the fuzzy c-means algorithm,[18] the final selected population from each run was examined for the presence of up to N/2 clusters, where N is the size of the selected population. Regardless of the ''fuzziness'' permitted within a cluster, values for the partition coefficient, entropy, and the nonfuzzy index invariably indicated total fuzzy partitioning. An absence of clustering was also indicated by the hierarchical average link method.[16] For example, from a population of 100 selected conformers, only one or two clusters (with two members) were formed below 1 Å, and none of these were formed below 0.5 Å. The remaining clusters were formed at well over 1 Å. The absence of clustering may be taken as an indication of the vastly complex conformation space that was searched.

Discussion

GAP1.0 EVALUATION OF [MET]-ENKEPHALIN

The μ- and δ-opioid receptors mediate the physiological effects of both [Leu]-enkephalin and the more predominant [Met]-enkephalin. Although evidence suggests that binding to these receptors leads to the opening of potassium channels, the physiological function of these receptors is unclear. Currently, it appears that δ receptors have been implicated in cardiovascular effects, and both δ and μ receptors are involved in pain modulation.[23] A more comprehensive understanding of the receptor structure–function relationship would benefit greatly from the elucidation of pharmacophoric elements within the enkephalin peptides. Although [Met]-enkephalin has been observed in a fully extended form in the crystalline state, the physiological significance of this structure—within the context of this molecule's tremendous conformational flexibility—is unclear.[24] Therefore, the characterization of conformational flexibility within [Met]-enkephalin is necessary to understand this ligand's bioactivity.

Unfortunately, even for small peptides, a systematic examination of the conformational space can tax current computational capabilities. For example, the conformation space of [Met]-enkephalin pentapeptide has been estimated to contain $\sim 10^{11}$ local minima.[6] Clearly, this situation will overwhelm a simplistic computational search. This has necessitated the development of alternative search strategies, which may employ a variation of the Monte Carlo approach, simulated annealing, or a molecular dynamics method. Many of these techniques have demonstrated their usefulness particularly with respect to finding minimum energy conformations that can be corroborated by comparison to experimentally elucidated structures. Search methods employing genetic algorithms provide an opportunity to gain novel conformational insight at a minimum cost of computer resources. Through a GA-driven search, it is possible to find many low energy conformers, each occurring at a different local potential minimum. Although it is reasonable to expect that a peptide's global minimum energy structure will be of biological importance, the significance of other energetically accessible conformations must not be discounted. Although the influence of molecular structure on bioactivity is still poorly understood, it is apparent that, within the receptor microenvironment, conformational flexibility strongly influences the molecule–receptor interaction. For example, theories of drug–receptor binding employ an ''induced fit'' model, in which the conformations of both the drug and the receptor site adopt complementary geometry. In peptide-based drug design, the inherent flexibility of the drug is altered not only to facilitate therapeutic receptor binding but also to minimize affinity to other ''toxic'' receptors and peptidases. Genetic algorithm-based search methods are useful in probing a molecule's conformation space for biologically significant structures and supplement other conformational search techniques. In this study, a number of genetic algorithm operators were assessed for their suitability in exploring the conformational diversity among low energy structures of [Met]-enkephalin. The lowest energy conformer found in this study clearly exhibited peptide backbone characteristics found in the global minimum energy structure reported by Nayeem et al.[6] (see Fig. 3a and b). Additionally, many diverse conformations were found to be within 5 to 10 kcal/mol of the lowest energy structure.

COMMENT ON GAP1.0 PERFORMANCE AND OTHER METHODS

The implicit parallelism of GAP1.0 permitted the rapid elucidation of thermally accessible regions of peptide conformational space. In this respect, GAP1.0 differs from previously reported conformational search methods. Modifications within the Monte Carlo method, such as the intro-

duction of biased sampling[25] or combination with energy minimization schemes,[6b] have shown great efficiency in finding global minimum energy structures. Similarly, simulated annealing[20] and direct search methods[26] have also shown success in this arena. More recently, the utility of GA-based search methods for global energy minimization tasks has been demonstrated across a broad spectrum of both hypothetical and realistic test cases.[14] For small organic molecules containing 12 or less rotatable bonds, Judson et al.[14c] found that, as the number of variable dihedral angles increases, GAs display greater CPU efficiency relative to other methods. For cyclic hexaglycine, the use of conformer subpopulations, or ''niches,'' was helpful in maintaining population diversity, which in turn improved the GA's effectiveness for global conformation search.[14d] For proteins, it has been suggested that the progress of a GA-based conformational space search is analogous to a real protein folding pathway.[14f] Dandekar and Argos have employed GAs to investigate the role of nonbonding interactions in simulations of hemerythrin,[14g] cytochromes b_{562}[14g] and c′,[14g] and a zinc finger motif.[14b] Clearly, GAs possess great potential for global conformational search tasks. The design of GAP1.0 acknowledges the significance of low energy conformational space regions which may be removed from the global energy minimum. In discovering many diverse [Met]-enkephalin conformers within slightly over 10 kcal/mol of the global minimum-energy structure, GAP1.0 indicated that this peptide may assume a wide variety of biologically significant conformations. Thus, the search strategy employed in GAP1.0 provides a novel emphasis on the biological relevance of regions—and not simply individual structures—in a conformational space.

GENETIC ALGORITHM OPERATORS IN GAP1.0

Crossover Operator

Genetic algorithms are distinguished from other so-called evolutionary programming paradigms by their use of the crossover operator. This operation permits structural data to be not only retained but also exploited for the purpose of discovering more useful information. By dictating the conditions under which crossover may occur, and by implementing this operator in a specific way, GA performance can be varied considerably. Many crossover methods previously investigated differ according to the number of crossover points used; in general, one-point crossover does not perform as well as two-point or multipoint crossover.[10] In this study, one-point crossover was inadequate in maintaining the chromosome population's diversity, resulting in preconvergence early within the GA run. The use of a two-point crossover scheme simply delayed preconvergence but did not prevent it. These results indicate the poor ability of these operators to generate new chromosomes when the population is nearing convergence. With only one or two crossover points, two similar parents are likely to exchange bit strings that are identical, yielding offspring that in turn are identical to the parents. Furthermore, for chromosomes which possess a high degree of epistasis—as is found within a binary string representation of a peptide conformation—one-point and two-point crossover operators are expected to perform poorly.[10] This is because schema of long defining length are likely to be disrupted during crossover. On the other hand, uniform crossover has been found to perform well despite epistasis. In this technique, the defining length of a schemata is irrelevant in regard to its likelihood of disruption during crossover. The probability of a schemata's disruption is dependent on its order; consequently, two schemata of the same order are equally likely to be disrupted even if one is ''spread'' over the entire chromosome and the other is contained between two closely spaced bit positions. Therefore, in addition to its capability to generate new chromosomes within highly converged populations, uniform crossover is also more successful in preserving good schema of long defining length. In this study, despite the relatively small population sizes used and the complexity of the peptide conformation space, the apparently greater robustness of uniform crossover drastically reduced the occurrence of preconvergence.

Mutation and Diversity Operators

In a simple genetic algorithm, the mutation operator possesses a purely exploratory function. As a population converges and new gene values arising from crossover occur less frequently, mutation becomes the dominant if not the sole exploratory mechanism. This observation has given rise to ''dynamic'' mutation strategies in which the mutation rate is varied over the course of the GA run.[10] For example, the mutation rate may be set high at the beginning when exploration has a good chance of improving population fitness, then lowered dur-

ing the middle of the GA run when the exploitation of good schema improves population fitness the most, and then raised again toward the end of the run when the population has become mostly converged. The limitation with this technique is that it introduces many new parameters which must be set at the beginning of the GA run: What mutation rates should be used? How should the mutation rate be varied? To circumvent these issues, in this study, the mutation rate was kept constant for the duration of each GA run. However, even with a high mutation rate, the exploration of conformational space was insufficient to suppress the preconvergent tendencies within the GA. Consequently, a second exploratory mechanism—the GEOMCHECK subroutine—was implemented to enforce diversity within the population by responding specifically to convergence. As such, the use of this ''diversity operator'' was greatly affected by the mutation rate. As the mutation rate was decreased, the increased convergence tendency invoked the GEOMCHECK subroutine to generate more mutants. Therefore, as the GA progressed, the nature of its exploratory capability shifted from that provided by mutation and crossover to that provided by the diversity operator. The introduction of this adaptive exploratory capability resulted in profound changes in GA performance. Chromosomes affected by the mutation operator have a good chance of retaining many characteristics of the parents, whereas those affected by the diversity operator are less likely to do so. Therefore, in GA runs with a low mutation rate, high usage of the diversity operator can give rise to many dissimilar groups of values for any given torsional angle. However, although these groups may be dissimilar with respect to each other, the values within a group are likely to be very similar due to the low mutation rate used. This was reflected in the scatterplots for the Gly residues of [Met]-enkephalin, in which many small, dense groups of points were supplanted by more diffuse scattering as the mutation rate was increased. The increased use of the diversity operator also affected the evolution of the average parent energy. At a low mutation rate, the average parent energy was influenced by members of the selected population taken from the large number of mutants that arose from the diversity operator. Therefore, quite often, the highest energy and the average energy of the parents were observed to increase from one generation to the next. However, the lowest energy in the population either remained the same or decreased, indicating that the capability of the GA to exploit the chromosomal information of the lowest energy conformer was not compromised by the use of the diversity operator.

Fitness and Selection Operators

The task of selecting conformers to form the next generation is based on a fitness value assigned to each member of the population. The first issue is deciding which of the conformer's characteristics will influence fitness. In this study, fitness was based solely on the conformer's energy. Alternatively, fitness can also be based on a combination of energy and geometry. This approach may be more appropriate when experimental data permit the derivation of conformational constraints, via NOEs from two-dimensional NMR studies, for example. By using both energy and geometry in assigning a conformer's fitness, the GA can be biased to explore a more restricted conformational space. Although this may reduce the complexity of the search problem, the conformers generated by this method may not accurately depict the conformational diversity present at low energy. For example, some small peptides have displayed rapid interconversion—on the NMR time scale—between conformations that were quite distinct.[27] Therefore, the NOEs of these peptides must be interpreted as ensemble averages which reflect all of the interconvertible conformation.[28] Failure to do so might lead a GA-driven search to generate conformations that resemble a single ''average structure'' rather than the different structures that collectively contribute to the observed NOE.

Once the fitness criteria have been established, actual fitness values must then be assigned to each conformer. In this study, the fitness-ranking method was used to insure that fitness values were evenly distributed between 0 and 1, regardless of the energy range within the population. This prevented the lowest energy conformers of any given generation from excessive mating. Alternatively, a constant fitness value may be assigned to every member of the parent population. This permits both low and high energy conformers to mate with equal probability, giving the crossover operator greater exploratory emphasis. In this case, the selection operator functions as the sole exploitative mechanism within the GA.

The selection of parents for mating is often closely linked to the replacement of parents by their offspring via the GA's fitness criteria, although not necessarily via the fitness values them-

selves. In this study, the fitness criterion—conformer energy—was used to select N/2 parent candidates from a total population of size N. The selected N/2 parent candidates were then assigned fitness values by the fitness-ranking method. This type of selection process, in which the parents and their offspring ''compete'' for selection, is known as ''steady-state replacement.'' Often, this results in only a few changes on going from one generation to the next. Another selection scheme—generational replacement—results in the replacement of a large percentage of the current generation by the offspring. The size of this percentage is determined by a preset parameter called the ''generation gap.'' Both methods have been shown to perform more or less equally.[10b,c] For this reason, the simpler steady-state replacement scheme was used in this study. Another simplification was made in the crossover operator. In most GA implementations, a conformer chosen for mating will have less than a 100% chance of mating successfully. For example, a conformer with a fitness value of 0.6 may have a 60% chance of being selected for mating. If it is chosen, then its chance of mating successfully may be preset at 90%. This has the effect of de-emphasizing the exploitative ability of the GA. In this study, conformers chosen for mating were given a 100% chance for successful mating. To insure that this simplification would not overemphasize exploitation, the second parent of the mating pair was selected randomly.

EXPLORATION, EXPLOITATION, AND GENETIC DRIFT IN GAP1.0

Theoretically, genetic algorithms can optimally balance the exploration and exploitation of a complex solution space. However, there are practical limitations which severely curtail the efficiency of the GA. For an investigation of peptide conformation space, the most serious of these obstacles are the interdependence of the molecule's torsional angles (epistasis) and the finite number of conformers that can exist within one generation. In this study, the uniform crossover operator was used to deal with gene interaction. Unfortunately, the use of a finite population size cannot be avoided and it is therefore necessary to consider the resulting genetic drift within the GA run. For optimization problems, the rate at which this phenomenon leads to the convergence of a gene imposes a minimum rate at which the GA must find the global optimum. For example, if genetic drift leads to the convergence of a gene within 200 generations, then the optimal solution should be found in less time than that. Solutions found in later generations may contain at least one converged gene which may not have an optimal value. In this study, the ability of the GA's exploitative mechanisms to improve fitness changed over the course of each run, whereas the rate of genetic drift did not. Therefore, at the start of a GA run the fast propagation of good schema led to a sharp decrease in conformer energy, outpacing the rate of genetic drift. As the GA progressed and the prevalence of good schema increased, crossover operations became less successful in improving the existing schema; thus, the GA's exploitative capability declined over time. Although it is possible to slow down genetic drift by using a high mutation rate, this also serves to debilitate the GA's exploitative ability through the disruption of good schema. This appears to present a dilemma: improved exploitative ability through a low mutation rate is offset by the increased rate of genetic drift. In this study, the exploratory capability afforded by the diversity operator was sufficient to counter this effect. Furthermore, the relative rates of genetic drift between runs using different mutation rates were reflected in the increased use of the diversity operator as each run progressed. From this it became clear that the mutation operator acted more as a restraint on the rate of genetic drift rather than as an effective exploratory mechanism. Exploration was mainly accomplished by the diversity operator while the crossover operator performed an effective exploitative role in the early stages of the GA run.

Conclusion

There is a need to expand the repertoire of conformational analysis techniques. GA-based conformational search methods present a twofold opportunity. In addition to providing another approach for global optimization tasks, GAs also offer a means of quickly determining the conformational possibilities that are available among biologically significant, low energy conformers. In this study's GA implementation, many diverse [Met]-enkephalin conformers were found to exist within a narrow, biologically feasible energy range. Also, the GA was successful in finding regions of the [Met]-enkephalin conformation space which encompassed the peptide backbone features of the global minimum energy structure previously re-

ported.[6,20] Many dissimilar low energy conformers were found within 10 kcal/mol of the global minimum energy structure, indicating that great conformational diversity is likely under physiological conditions.

Genetic algorithm operations permit great adaptability in searching peptide conformation space. The uniform crossover operator was robust in dealing with the inherent epistasis of peptide folding. The mutation operator's influence on genetic drift consequently affected the diversity operator's role as the main exploratory mechanism within the GA. By choosing the correct parameter settings, the GA was tailored to fulfill different tasks. For example, a low mutation rate generally led to the discovery of more stable lowest energy conformers, whereas a high mutation rate resulted in a more homogeneous distribution of phi and psi angles over the parent population. Also, by combining the scatterplots from runs using different initial populations (but the same parameters), it was possible to gain a qualitative idea of the conformational flexibility within the peptide backbone. As population size was increased, the effects of genetic drift became ''diluted.'' The lowest average parent conformer energy was generally found at mutation rate 0.05 for a population size of 100, and at either 0.03 or 0.05 for a population size of 200. It was of interest to note that population size had no effect on the GA's ability to find the most stable low energy conformers; often, the more stable conformers were found among populations of 50 parent conformers.

Alternative strategies for both exploration and exploitation may improve adaption within the genetic algorithm. Recent examples of this include niching[14d] and diploidy.[10] There is also a need to develop exploitative operators that are more effective than crossover during the later stages of a GA run. Additionally, the hybridization of energy minimization routines with the genetic algorithm should enhance exploitation without promoting genetic drift.

Acknowledgments

D. F. W. acknowledges the support of an Ontario Ministry of Health Career Scientist Award. Dr. Heather L. Gordon (Neurochem Inc.) kindly provided fuzzy cluster analysis programs and insight into their use. Finally, the authors thank Mark N. Anderson (Andyne Computing Ltd.) for many useful discussions throughout the course of this study.

Source code for GAP1.0 is available on request.

References

1. W. M. Pardridge, *Peptide Drug Delivery to the Brain*, Raven Press, New York, 1991, p. 1.
2. R. B. Silverman, *The Organic Chemisty of Drug Design and Drug Action*, Academic Press, San Diego, CA, 1992, p. 52.
3. (a) N. L. Allinger, *J. Am. Chem. Soc.*, **99**, 8127 (1977); (b) R. B. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.*, **4**, 187 (1983); (c) S. L. Mayo, B. D. Olafson, and W. A. Goddard III, *J. Phys. Chem.*, **94**, 8897 (1990); (d) F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361 (1975); (e) P. K. Weiner and P. A. Kollman, *J. Comput. Chem.*, **2**, 287 (1981).
4. H. A. Scheraga, In *Reviews in Computational Chemistry, Vol. 3*, K. B. Lipkowitz and D. B. Boyd, Eds., VCH, New York, 1992, p. 73.
5. (a) N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Phys. Chem.*, **21**, 1087 (1953); (b) M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, New York, 1992, p. 110.
6. (a) Z. Li and H. A. Scheraga, *J. Mol. Struct. (Theochem)*, **179**, 333 (1988); (b) A. Nayeem, J. Vila, and H. A. Scheraga, *J. Comput. Chem.*, **12**, 594 (1991).
7. (a) W. F. van Gunsteren and H. J. C. Berendsen, *Angew. Chem. Int. Ed. Engl.*, **29**, 992 (1990); (b) M. Karplus, *Israel J. Chem.*, **27**, 121 (1986) (c) J. M. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley, New York, 1992, p. 38.
8. (a) A. T. Brünger, J. Kuriyan, and M. Karplus, *Science*, **235**, 458 (1987); (b) A. T. Brünger and M. Karplus, *Acc. Chem. Res.*, **24**, 54 (1991).
9. J. Heitkotter and D. Beasley, Eds. *The Hitchhiker's Guide to Evolutionary Computation: A List of Frequently Asked Questions (FAQ)*, 1994. Available via anonymous ftp from rtfm.mit.edu:/pub/usenet/news.answers/ai-faq/genetic.
10. (a) L. Davis, *Genetic Algorithms and Simulated Annealing*. Pitman, London, 1987; (b) D. Beasley, D. R. Bull, and R. R. Martin, *University Computing*, **15**, 58 (1993); (c) D. Beasley, D. R. Bull, and R. R. Martin, *Univ. Comput.*, **15**, 170 (1993).
11. J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, Cambridge, MA, 1992.
12. C.B. Lucasius and G. Kateman, *Trends Anal. Chem.*, **10**, 254 (1991).
13. (a) R. S. Judson, E. P. Jaeger, and A. M. Treasurywala, *J. Mol. Struc. (Theochem)*, **308**, 191 (1994); (b) D. E. Walters and R. M. Hinds, *J. Med. Chem.*, **37**, 2527 (1994); (c) C. M. Oshiro, I. D. Kuntz, and J. S. Dixon, *J. Comput.-Aided. Mol. Design*, **9**, 113 (1995).
14. (a) R. S. Judson, M. E. Colvin, J. C. Meza, A. Huffer, and D. Gutierrez, *Int. J. Quantum Chem.*, **44**, 277 (1992); (b) T. Dandekar and P. Argos, *Protein Eng.*, **5**, 637 (1992); (c) R. S. Judson, E. P. Jaeger, A. M. Treasurywala, and M. L. Peter-

son, *J. Comput. Chem.*, **14**, 1407 (1993); (d) D. B. McGarrah and R. S. Judson, *J. Comput. Chem.*, **14**, 1385 (1993); (e) P. Tufféry, C. Etchebest, S. Hazout, and R. Lavery, *J. Comput. Chem.*, **14**, 790 (1993); (f) R. Unger and J. Moult, *J. Mol. Biol.*, **231**, 75 (1993); (g) T. Dandekar and P. Argos, *J. Mol. Biol.*, **236**, 844 (1994); (h) J. Mestres and G. E. Scuseria, *J. Comput. Chem.*, **16**, 729 (1995).

15. *ECEPP: Empirical Conformation Energy Program for Peptides*, Cornell University, Ithaca, NY, (QCPE Program No. 454).
16. (a) M. S. Aldenferer and R. K. Blashfield, *Cluster Analysis*, Sage, Beverly Hills, CA, 1984; (b) W. Vogt and D. Nagel, *Clin. Chem.*, **38**, 182 (1992); (c) A. E. Torda and W. F. van Gunsteren, *J. Comput. Chem.*, **15**, 1331 (1994); (d) P. S. Shenkin and D. Q. McDonald, *J. Comput. Chem.*, **15**, 899 (1994).
17. *S-PLUS*, MathSoft Inc., Seattle, WA.
18. (a) J. C. Bezdek, R. Ehrlich, and W. Full, *Comput. Geosci.*, **10**, 191 (1984); (b) H. L. Gordon and R. L. Somorjai, *Prot. Struct. Function Genet.*, **14**, 249 (1992).
19. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran, 2nd Ed.*, Cambridge University Press, Cambridge, 1992, pp. 393, 397, 411.
20. (a) S. R. Wilson and W. Cui, *Biopolymers*, **29**, 225 (1990); (b) H. Kawai, T. Kikuchi, and Y. Okamoto, *Protein Eng.*, **3**, 85 (1989).
21. (a) G. N. Ramachandran and V. Sasisekharan, *Adv. Prot. Chem.*, **23**, 283 (1968); (b) M. Levitt, *J. Mol. Biol.*, **104**, 59 (1976).
22. G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure*, Springer, New York 1979 p. 17.
23. E. J. Simon and J. M. Hiller, In *Basic Neurochemistry, 5th Ed.*, G. J. Siegel, B. W. Agranoff, R. W. Albers, and P. B. Molinoff, Eds., Raven Press, New York, 1994, p. 335.
24. J. R. Deschamps, C. George, and J. L. Flippen-Anderson, *Biopolymers* (*Peptide Science*), **40**, 121 (1996).
25. J. K. Shin and M. S. Jhon, *Biopolymers*, **31**, 177 (1991).
26. J. C. Meza and M. L. Martinez, *J. Comput. Chem.*, **15**, 627 (1994).
27. H. Pepermans, D. Tourwé, G. van Binst, R. Boelens R. M. Scheek, W. F. van Gunsteren, and R. Kaptein, *Biopolymers*, **27**, 323 (1988).
28. A. E. Torda and W. F. van Gunsteren, In *Reviews in Computational Chemistry*, Vol. 3, K. B. Lipkowitz and D. B. Boyd, Eds., VCH, New York, 1992, p. 143.